HITIQA: An Interactive Question Answering System A Preliminary Report

Sharon Small, Ting Liu, Nobuyuki Shimizu, and Tomek Strzalkowski

ILS Institute
The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
{small,tl7612,ns3203,tomek}@albany.edu

Abstract

HITIQA is an interactive question answering technology designed to allow intelligence analysts and other users of information systems to pose questions in natural language and obtain relevant answers, or the assistance they require in order to perform their tasks. Our objective in HITIQA is to allow the user to submit exploratory, analytical, non-factual questions, such as *"What has been Russia's reaction to U.S. bombing of Kosovo?"* The distinguishing property of such questions is that one cannot generally anticipate what might constitute the answer. While certain types of things may be expected (e.g., diplomatic statements), the answer is heavily conditioned by what information is in fact available on the topic. From a practical viewpoint, analytical questions are often under-specified, thus casting a broad net on a space of possible answers. Therefore, clarification dialogue is often needed to negotiate with the user the exact scope and intent of the question.

1 Introduction

HITIQA project is part of the ARDA AQUAINT program that aims to make significant advances in the state of the art of automated question answering. In this paper we focus on two aspects of our work:

1. Question Semantics: how the system "understands" user requests.
2. Human-Computer Dialogue: how the user and the system negotiate this understanding.

We will also discuss very preliminary evaluation results from a series of pilot tests of the system conducted by intelligence analysts via a remote internet link.

2 Factual vs. Analytical

The objective in HITIQA is to allow the user to submit and obtain answers to exploratory, analytical, non-factual questions. There are very significant differences between factual, or fact-finding, and analytical question answering. A factual question seeks pieces of information that would make a corresponding statement true (i.e., they become facts): "How many states are in the U.S.?" / "There are X states in the U.S." In this sense, a factual question usually has just one correct answer that can generally, be judged for its truthfulness. By contrast, an analytical question is when the "truth" of the answer is more a matter of opinion and may depend upon the context in which the question is asked. Answers to analytical questions are rarely unilateral, indeed, a mere "correct" answer may have limited value, and in some cases may not even be determinate ("Which college is the best?", "How do I stop my baby's crying?"). Instead, answers to analytical questions are often judged as helpful, or useful, or satisfactory, etc. "Technically correct" answers (e.g., "feed the baby milk") may be considered as irrelevant or at best unresponsive.

The distinction between factual and analytical questions depends primarily on the intention of the person who is asking, however, the form of a question is often indicative of which of the two classes it is more likely to belong to. Factual questions can be classified into a number of syntactic formats ("question typology") that aids in automatic processing.

Factual questions display a fairly distinctive "answer type", which is the type of the information piece needed to fulfill the statement. Recent automated systems for answering factual questions deduct this expected answer type from the form of the question and a finite list of possible answer

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
HITIQA: An Interactivev Question Answering System. A Preliminary Report	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The State University of New York at Albany,1400 Washington Avenue,Albany,NY,12222**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

types. For example, "Who was the first man in space" expects a "person" as the answer, while "How long was the Titanic?" expects some length measure as an answer, probably in yards and feet, or meters. This is generally a very good strategy, that has been exploited successfully in a number of automated QA systems that appeared in recent years, especially in the context of TREC QA[1] evaluations (Harabagiu et al., 2000; Hovy et al., 2000; Prager at al., 2001).

This process is not easily applied to analytical questions. This is because the type of an answer for analytical questions cannot always be anticipated due to their inherently exploratory character. In contrast to a factual question, an analytical question has an unlimited variety of syntactic forms with only a loose connection between their syntax and the expected answer. Given the unlimited potential of the formation of analytical questions, it would be counter-productive to restrict them to a limited number of question/answer types. Even finding a non-strictly factual answer to an otherwise simple question about Titanic length (e.g., "two football fields") would push the limits of the answer-typing approach. Therefore, the formation of an answer should instead be guided by the topics the user is interested in, as recognized in the query and/or through the interactive dialogue, rather than by a single type as inferred from the query in a factual system.

This paper argues that the semantics of an analytical question is more likely to be deduced from the information that is considered relevant to the question than through a detailed analysis of their particular form. While this may sound circular, it needs not be. Determining "relevant" information is not the same as finding an answer; indeed we can use relatively simple information retrieval methods (keyword matching, etc.) to obtain perhaps 50 or 100 "relevant" documents from a database. This gives us an initial answer space to work on in order to determine the scope and complexity of the answer. In our project, we use structured templates, which we call *frames* to map out the content of pre-retrieved documents, and subsequently to delineate the possible meaning of the question (Section 6).

[1] TREC QA is the annual Question Answering evaluation sponsored by the U.S. National Institute of Standards and Technology www.trec.nist.gov.

3 Document Retrieval

When the user poses a question to a system sitting atop a huge database of unstructured data (text files), the first order of business is to reduce that pile to perhaps a handful of documents where the answer is likely to be found. This means, most often, document retrieval, using fast but non-exact selection methods. Questions are tokenized and sent to a document retrieval engine, such as Smart (Buckley, 1985) or InQuery (Callan et al., 1992). Noun phrases and verb phrases are extracted from the question to give us a list of potential topics that the user may be interested in.

In the experiments with the HITIQA prototype, see Figure 1, we are retrieving the top fifty documents from three gigabytes of newswire (AQUAINT corpus plus web-harvested documents).

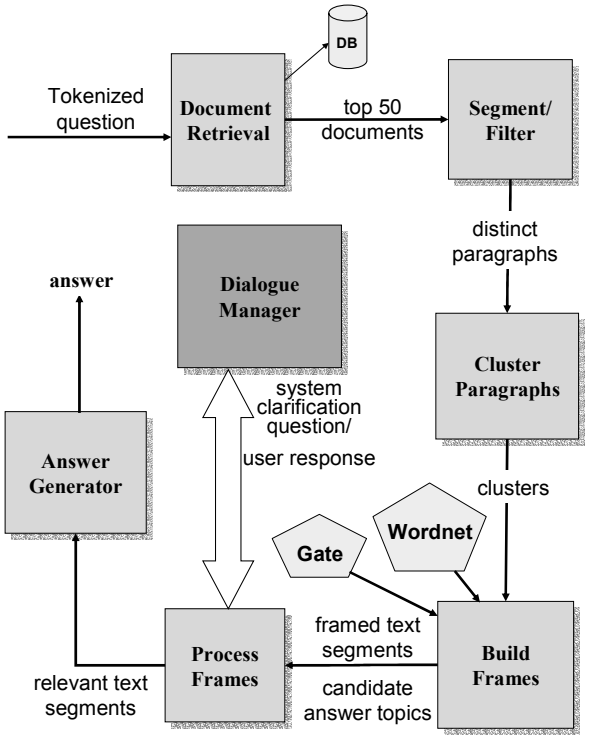

Figure 1: HITIQA preliminary architecture

4 Data Driven Semantics of Questions

The set of documents and text passages returned from the initial search is not just a random subset of the database. Depending upon the quality (recall and precision) of the text retrieval system avail-

able, this set can be considered as a first stab at understanding the user's question by the machine. Again, given the available resources, this is the best the system can do under the circumstances. Therefore, we may as well consider this collection of retrieved texts (*the Retrieved Set*) as the meaning of the question as understood by the system. This is a fair assessment: the better our search capabilities, the closer this set would be to what the user may accept as an answer to the question.

We can do better, however. We can perform automatic analysis of the retrieved set, attempting to uncover if it is a fairly homogenous bunch (i.e., all texts have very similar content), or whether there are a number of diverse topics represented there, somehow tied together by a common thread. In the former case, we may be reasonably confident that we have the answer, modulo the retrievable information. In the latter case, we know that the question is more complex than the user may have intended, and a negotiation process is needed.

We can do better still. We can measure how well each of the topical groups within the retrieved set is "matching up" against the question. This is accomplished through a framing process described later in this paper. The outcome of the framing process is twofold: firstly, the alternative interpretations of the question are ranked within 3 broad categories: *on-target, near-misses and outliers*. Secondly, salient concepts and attributes for each topical group are extracted into topic frames. This enables the system to conduct a meaningful dialogue with the user, a dialogue which is wholly content oriented, and thus entirely data driven.

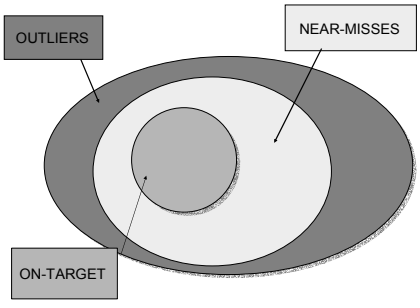

Figure 2: Answer Space Topology. The goal of interactive QA it to optimize the ON-TARGET middle zone.

5 Clustering

We use n-gram-based clustering of text passages and concept extraction to uncover the main topics, themes and entities in this set.

Retrieved documents are first broken into naturally occurring paragraphs. Duplicate paragraphs are filtered out and the remaining passages are clustered using a combination of hierarchical clustering and n-bin classification (details of the clustering algorithm can be found in Hardy et al., 2002a). Typically three to six clusters are generated out of the top 50 documents, which may yield as many as 1000 passages. Each cluster represents a topic theme within the retrieved set: usually an alternative or complimentary interpretation of the user's question.

A list of topic labels is assigned to each cluster. A topic label may come from one of two places: First, the texts in the cluster are compared against the list of key phrases extracted from the user's query. For each match found, the matching phrase is used as a topic label for the cluster. If a match with the key phrases from the question cannot be obtained, Wordnet is consulted to see if a common ancestor can be found. For example, "rifle" and "machine gun" are kinds of "weaponry" in Wordnet, which allows an indirect match between a question about weapon inspectors and a text reporting a discovery by the authorities of a cache of "rifles" and "machine guns".

6 Framing

In HITIQA we use a *text framing* technique to delineate the gap between the meaning of the user's question and the system "understanding" of this question. The framing is an attempt to impose a partial structure on the text that would allow the system to systematically compare different text pieces against each other and against the question, and also to communicate with the user about this. In particular, the framing process may uncover topics and themes within the retrieved set which the user has not explicitly asked for, and thus may be unaware of their existence. Nonetheless these may carry important information – the NEAR-MISSES in Figure 2.

In the current version of the system, frames are fairly generic templates, consisting of a small number of attributes, such as LOCATION, PERSON, COUNTRY, ORGANIZATION, etc. Future versions of HITIQA will add domain specialized frames, for example, we are currently constructing frames for the Weapons Non-proliferation Domain. Most of the frame attributes are defined in advance, how-

ever, dynamic frame expansion is also possible. Each of the attributes in a frame is equipped with an extractor function which specializes in locating and extracting instances of this attribute in the running text. The extractors are implemented using information extraction utilities which form the kernel of Sheffield's GATE[2] system. We have modified GATE to separate organizations into companies and other organizations, and we have also expanded by adding new concepts such as industries. Therefore, the framing process resembles strongly the template filling task in information extraction (cf. MUC[3] evaluations), with one significant exception: while the MUC task was to fill in a template using potentially any amount of source text (Humphreys et al., 1998), the framing is essentially an inverse process. In framing, potentially multiple frames can be associated with a small chunk of text (a passage or a short paragraph). Furthermore, this chunk of text is part of a cluster of very similar text chunks that further reinforce some of the most salient features of these texts. This makes the frame filling a significantly less error-prone task – our experience has been far more positive than the MUC evaluation results may indicate. This is because, rather than trying to find the most appropriate values for attributes from among many potential candidates, we in essence fit the frames over small passages[4].

Therefore, data frames are built from the retrieved data, after clustering it into several topical groups. Since clusters are built out of small text passages, we associate a frame with each passage that serves as a seed of a cluster. We subsequently merge passages, and their associated frames whenever anaphoric and other cohesive links are detected.

A very similar process is applied to the user's question, resulting in a *Goal Frame* which can be subsequently compared to the data frames obtained from retrieved data. For example, the Goal Frame generated from the question, "*How has pollution in the Black Sea affected the fishing industry, and what are the sources of this pollution?*" is shown in Figure 3 below.

TOPIC*:[pollution, industry, sources]*
LOCATION: *[Black Sea]*
INDUSTRY*:[fishing]*

Figure 3: HITIQA generated Goal Frame

TOPIC: *pollution*
SUB-TOPIC: *[sources]*
LOCATION: *[Black Sea]*
INDUSTRY :*[fisheries, tourism]*
TEXT: *[In a period of only three decades (1960's-1980's), the **Black Sea** has suffered the catastrophic degradation of a major part of its natural resources. Particularly acute problems have arisen as a result of **pollution** (notably from nutrients, fecal material, solid waste and oil), a catastrophic decline in commercial fish stocks, a severe decrease in **tourism** and an uncoordinated approach towards coastal zone management. Increased loads of nutrients from rivers and coastal **sources** caused an overproduction of phytoplankton leading to extensive eutrophication and often extremely low dissolved oxygen concentrations. The entire ecosystem began to collapse. This problem, coupled with **pollution** and irrational exploitation of fish stocks, started a sharp decline in **fisheries** resources.]*
RELEVANCE: *Matches on all elements found in goalframe*

Figure 4: A HITIQA generated data frame. Words in bold were used to fill the Frame.

The data frames are then compared to the Goal Frame. We pay particular attention to matching the topic attributes, before any other attributes are considered. If there is an exact match between a Goal Frame topic and the text being used to build the data frame, then this becomes the data frame's topic as well. If more than one match is found, the subsequent matches become the sub-topics of the data frame. On the other hand, if no match is possible against the Goal Frame topic, we choose the topic from the list of the Wordnet generated hypernyms. An example data frame generated from the text retrieved in response to the query about the Black Sea is shown in Figure 4. After the initial framing is done, frames judged to be related to the same concept or event, are merged together and values of their attributes are combined.

7 Judging Frame Relevance

We judge a particular data frame as relevant, and subsequently the corresponding segment of text as relevant, by comparison to the Goal Frame. The

[2] GATE is Generalized Architecture for Text Engineering, an information extraction system developed at the University of Sheffield (Cunningham, 2000).
[3] MUC, the Message Understanding Conference, funded by ARPA, involved the evaluation of information extraction systems applied to a common task.
[4] We should note that selecting the right frame type for a passage is an important pre-condition to "understanding".

data frames are scored based on the number of conflicts found between them and the Goal Frame. The conflicts are mismatches on values of corresponding attributes. If a data frame is found to have no conflicts, it is given the highest relevance rank, and a conflict score of zero. All other data frames are scored with an incrementing conflict value, one for frames with one conflict with the Goal Frame, two for two conflicts etc. Frames that conflict with all information found in the query are given a score of 99 indicating the lowest relevancy rank. Currently, frames with a conflict score of 99 are excluded from further processing. The frame in Figure 4 is scored as fully relevant to the question (0 conflicts).

8 Enabling Dialogue with the User

Framed information allows HITIQA to automatically judge some text as relevant and to conduct a meaningful dialogue with the user as needed on other text. The purpose of the dialogue is to help the user to navigate the answer space and to solicit from the user more details as to what information he or she is seeking. The main principle here is that the dialogue is at the information semantic level, not at the information organization level. Thus, it is okay to ask the user whether information about the AIDS conference in Cape Town should be included in the answer to a question about combating AIDS in Africa. However, the user should never be asked if a particular keyword is useful or not, or if a document is relevant or not. We have developed a 3-pronged strategy:

1. Narrowing dialogue: ask questions that would allow the system to reduce the size of the answer set.
2. Expanding dialogue: ask questions that would allow the system to decide if the answer set needs to be expanded by information just outside of it (near-misses).
3. Fact seeking dialogue: allow the user to ask questions seeking additional facts and specific examples, or similar situations.

Of the above, we have thus far implemented the first two options as part of the preliminary clarification dialogue. The clarification dialogue is when the user and the system negotiate the task that needs to be performed. We can call this a "triaging stage", as opposed to the actual problem solving stage (point 3 above). In practice, these two stages are not necessarily separated and may be overlapping throughout the entire interaction. Nonetheless, these two have decidedly distinct character and require different dialogue strategies on the part of the system.

Our approach to dialogue in HITIQA is modeled to some degree upon the mixed-initiative dialogue management adopted in the AMITIES project (Hardy et al., 2002b). The main advantage of the AMITIES model is its reliance on data-driven semantics which allows for spontaneous and mixed initiative dialogue to occur.

By contrast, the major approaches to implementation of dialogue systems to date rely on systems of functional transitions that make the resulting system much less flexible. In the grammar-based approach, which is prevalent in commercial systems, such as in various telephony products, as well as in practically oriented research prototypes[5], (e.g., DARPA, 2002; Seneff and Polifoni, 2000; Ferguson and Allen, 1998) a complete dialogue transition graph is designed to guide the conversation and predict user responses, which is suitable for closed domains only. In the statistical variation of this approach, a transition graph is derived from a large body of annotated conversations (e.g., Walker, 2000; Litman and Pan, 2002). This latter approach is facilitated through a dialogue annotation process, e.g., using Dialogue Act Markup in Several Layers (DAMSL) (Allen and Core, 1997), which is a system of functional dialogue acts.

Nonetheless, an efficient, spontaneous dialogue cannot be designed on a purely functional layer. Therefore, here we are primarily interested in the semantic layer, that is, the information exchange and information building effects of a conversation. In order to properly understand a dialogue, both semantic and functional layers need to be considered. In this paper we are concentrating exclusively on the semantic layer.

9 Clarification Dialogue

Data frames with a conflict score of zero form the initial kernel answer space. Depending upon the size of this set and the presence of other frames outside of it, the system either proceeds to generate the answer or initiates a dialogue with the user. For

[5] A notable exception is CU Communicator developed at University of Colorado (Ward and Pellom, 1999)

example, if the answer space appears too large or varied, e.g. consists of many different topics, the system may ask the user how to narrow it. Alternatively, the presence of large groups of texts frames with near-miss frames assigned to them (i.e., frames with 1 or 2 conflicts with the Goal Frame) may indicate that the answer space is actually larger, and the user will be consulted about a possible broadening of the question. Currently, we only initiate a clarification dialogue for 1-conflict frames.

A 1-conflict frame has only a single attribute mismatch with the Goal Frame. This could be a mismatch on any attribute, for example, LOCATION, or ORGANIZATION, or TIME, etc. A special case arises when the conflict occurs on the TOPIC attribute. Since all other attributes match, we may be looking at potentially different events or situations involving the same entities, or occurring at the same location or time. The purpose of the clarification dialogue in this case is to probe which of these topics may be of interest to the user. This is illustrated in the exchange below recorded during an evaluation session with an intelligence analyst:

User: *"Who is Elizardo Sanchez?"*
HITIQA*: "Are you interested in seeing information about civil rights as it is related to Elizardo Sanchez?*

ONE-CONFLICT FRAME
TOPIC**: *civil_rights***
SUB-TOPIC*: []*
LOCATION: *[Cuba, Cuba, Cuba]*
COMPANIES: *[]*
PEOPLE: *[Sanchez, Sanchez]*
DOCUMENT DATE: *[2000, 1, 11]*
SOURCE*: HAVANA (AP) –*
TEXT: *[``I consider that the situation for civil and political rights in **Cuba** has worsened over the **past year**... owing to that **Cuba** continues to be the only closed society in this hemisphere," **Sanchez** said. ``There have been no significant release of prisoners, the number of people sanctioned or processed for political motives increased. **Sanchez**, who himself spent many years in **Cuban** prisons, is among the communist island's best known opposition activists. The commission he heads issues a report on civil rights every six months, along with a list of people it considers to be imprisoned for political motives.]*

Figure 5: One of the Frames that were used in generating Sanchez dialogue. Words in bold were used to fill the Frame.

In order to understand what happened here, we need to note first that the Goal Frame for the user question does not have any specific value assigned to its TOPIC attribute. This of course is as we would expect it: the question does not give us a hint as to what information we need to look for or may be hoping to find about Sanchez. This also means that all the text frames obtained from the retrieved set for this question will have at least one conflict, near-misses. One such text frame is shown in Figure 5: its topic is "civil rights" and it about Sanchez. HITIQA thus asks if "civil rights" is a topic of interest to the user. If the user responds positively, this topic will be added to the answer space.

The above dialogue strategy is applicable to other attribute mismatch cases, and produces intelligent-sounding responses from the system. During the dialogue, as new information is obtained from the user, the Goal Frame is updated and the scores of all the data frames are reevaluated. The system may interpret the new information as a positive or negative. Positives are added to the Goal Frame. Negatives are stored in a Negative-Goal Frame and will also be used in the re-scoring of the data frames, possibly causing conflict scores to increase. The Negative-Goal Frame is created when HITIQA receives a negative response from the user. The Negative-Goal Frame includes information that HITIQA has identified as being of no interest to the user. If the user responds the equivalent of "yes" to the system clarification question in the Sanchez dialogue, *civil_rights* will be added to the topic list in the Goal Frame and all one-conflict frames with a civil_rights topic will be re-scored to Zero conflicts, two-conflict frames with civil_rights as a topic will be rescored to one, etc. If the user responds "no", the Negative-Goal Frame will be generated and all frames with civil_rights as a topic will be rescored to 99 in order to remove them from further processing.

The clarification dialogue will continue on the topic level until all the significant sets of NEAR-MISS frames are either included in the answer space (through user broadening the scope of the question that removes the initial conflicts) or dismissed as not relevant. When HITIQA reaches this point it will re-evaluate the data frames in its answer space. If there are too many answer frames now (more than a pre-determined upper threshold), the dialogue manager will offer to the user to narrow the question using another frame attribute. If the size of the new answer space is still too small (i.e., there are many unresolved near-miss frames),

the dialogue manager will suggest to the user ways of further broadening the question, thus making more data frames relevant, or possibly retrieving new documents by adding terms acquired through the clarification dialogue. When the number of frames is within the acceptable range, HITIQA will generate the answer using the text from the frames in the current answer space. The user may end the dialogue at any point and have an answer generated given the current state of the frames.

9.1 Narrowing Dialogue

HITIQA attempts to reduce the number of frames judged to be relevant through a Narrowing Dialogue. This is done when the answer space contains too many elements to form a succinct answer. This typically happens when the initial question turns out to be too vague or unspecific, with respect to the available data.

9.2 Broadening Dialogue

As explained before, the system may attempt to increase the number of frames judged relevant through a Broadening Dialogue (BD), whenever the answer space appears too narrow, i.e., contains too few zero-conflict frames. We are conducting further experiments to define this situation more precisely. Currently, the BD will only occur if there are one-conflict frames, or near misses. Broadening questions can be asked about any of the attributes which have values in the Goal Frame.

10 Answer Generation

Currently, the answer is simply composed of text passages from the zero conflict frames. The text of these frames are ordered by date and outputted to the user. Typically the answer to these analytical type questions will require many pages of information. Example 1 below shows the first portion of the answer generated by HITIQA for the Black Sea query. Current work is focusing on answer generation.

2002:
The Black Sea is widely recognized as one of the regional seas most damaged by human activity. Almost one third of the entire land area of continental Europe drains into this sea… major European rivers, the Danube, Dnieper and Don, discharge into this sea while its only connection to the world's oceans is the narrow Bosphorus Strait. The Bosphorus is as little as 70 meters deep and 700 meters wide but the depth of the Black Sea itself exceeds two kilometers in places. Contaminants and nutrients enter the Black Sea via river run-off mainly and by direct discharge from land-based sources. The management of the Black Sea itself is the shared responsibility of the six coastal countries: Bulgaria, Georgia, Romania, Russian Federation, Turkey, and Ukraine…

Example 1: Partial answer generated by HITIQA to the Black Sea query.

11 Evaluations

We have just completed the first round of a pilot evaluation for testing the interactive dialogue component of HITIQA. The purpose of this first stage of evaluation is to determine what kind of dialogue is acceptable/tolerable to the user and whether an efficient navigation though the answer space is possible. HITIQA was blindly tested by two different analysts on eleven different topics. Five different groups participated, but no analyst tested more than one system, as system comparison was not a goal. The analysts were given complete freedom in forming their queries and responses to HITIQA's questions. They were only provided with descriptions of the eleven topics the systems would be tested on. The analysts were given 15 minutes for each topic to arrive at what they believed to be an acceptable answer. During testing a Wizard (human) was allowed to intervene if HITIQA generated a dialogue question/response that was felt inappropriate. The Wizard was able to override the system and send a Wizard generated question/response to the analyst. The HITIQA Wizard intervened an average of 13% of the time.

These results are for information purposes only as it was not a formal evaluation. HITIQA earned an average score of 5.8 from both Analysts for dialogue, where 1 was "extremely dissatisfied" and 7 was "completely satisfied". The highest score possible was a 7 for each dialogue. The Analysts were asked to grade each scenario for success or failure. We divide the failures from both analysts into three categories:

1) the user gives up on the system for the given scenario(9%)
2) the 15 minute time limit was up(13%)
3) the data was not in the database(9%)

HITIQA had a 63% success rate for Analyst 1 and a 73% success rate for Analyst 2. It is unclear how

these results should be interpreted, if at all, as the evaluation was a mere pilot, mostly to test the mechanics of the setup. We know only that a human Wizard equipped with all necessary information can easily achieve 100% success in this test. What is still needed is a baseline performance, perhaps based on using an ordinary keyword-based search engine.

12 Future Work

This paper describes a work in progress. We expect that the initial specification of content frame will evolve as we subject the initial system to more demanding evaluations. Currently, the frames are not topically specialized, and this appears the most logical next refinement, i.e., develop several (10-30) types of frames covering different classes of events, from politics to medicine to science to international economics, etc. This is expected to increase the accuracy of the dialogue as is the interactive visualization which is also under development. Answer generation will involve fusion of information on the frame level, and is currently in an initial phase of implementation.

Acknowledgements

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000.

References

J. Allen. and Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. http://www.cs.rochester.edu/research/cisd/ resources/damsl/

Bagga, A., T. Strzalkowski, and G.B. Wise. 2000. PartsID: A Dialog-Based System for Identifying Parts for Medical Systems. *Proc. of the ANLP-NAACL-2.*

Chris Buckley. May 1985. Implementation of the Smart information retrieval system. *Technical Report TR85-686*, Department of Computer Science, Cornell University, Ithaca, NY.

James P. Callan, W. Bruce Croft, Stephen M. Harding 1992. The INQUERY Retrieval System. *Proc. of DEXA-92, 3rd International Conference on Database and Expert Systems Applications.* 78-83.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan and Y. Wilks. 2000 Experience of using GATE for NLP R&D. *In Coling 2000 Workshop on Using Toolsets and Architectures To Build NLP Systems.*

DARPA Communicator Program. 2002. http://www.darpa.mil/iao/communicator

Grinstein, G.G., Levkowitz, H., Pickett, R.M., Smith, S. 1993. "Visualization alternatives: non-pixel based images," *Proc. of IS&T 46th Annual Conf.* 132-133.

George Ferguson and James Allen. 1998. "TRIPS: An Intelligent Integrated Problem-Solving Assistant," in *Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI. 567-573.

H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, B. Wise and X. Zhang 2002a. Cross-Document Summarization by Concept Classification. *Proceedings of SIGIR-2002*, Tampere, Finland.

H. Hardy, K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu and N. Webb. 2002b. Multi-layer Dialogue Annotation for Automated Multilingual Customer Service. *ISLE Workshop*, Edinburgh, Scotland.

Harabagiu, S., M. Pasca and S. Maiorano. 2000. Experiments with Open-Domain Textual Question Answering. In *Proc. of COLING-2000*. 292-298.

Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks. 1998. Description of the LaSIE-II System as Used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7.)*

Judith Hochberg, Nanda Kambhatla and Salim Roukos. 2002. A Flexible Framework for Developing Mixed-Initiative Dialog Systems. *Proc. of 3rd SIGDIAL* Workshop on Discourse and Dialogue, Philadelphia.

Hovy, E., L. Gerber, U. Hermjakob, M. Junk, C-Y. Lin. 2000. Question Answering in Webclopedia. *Notebook Proceedings of Text Retrieval Conference (TREC-9).*

Johnston, M., Ehlen, P., Bangalore, S., Walker., M., Stent, A., Maloor, P., and Whittaker, S. 2002. MATCH: An Architecture for Multimodal Dialogue Systems. In *Meeting of the Association for Computational Linguistics* , 2002.

Diane J. Litman and Shimei Pan. Designing and Evaluating an Adaptive Spoken Dialogue System. 2002. *User Modeling and User-Adapted Interaction.* 12(2/3):111-137.

Miller, G.A. 1995. WordNet: A Lexical Database. *Comm. of the ACM*, 38(11):39-41.

John Prager, Dragomir R. Radev, and Krzysztof Czuba. Answering what-is questions by virtual annotation. In Human Language Technology Conference, Demonstrations Section, San Diego, CA, 2001.

S. Seneff and J. Polifroni, ``Dialogue Management in the MERCURY Flight Reservation System," *Proc. ANLP-NAACL 2000, Satellite Workshop*, 1-6, Seattle, WA, 2000.

Marilyn A. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email . *Journal of Artificial Intelligence Research*.12:387-416.

W. Ward and B. Pellom. 1999. The CU Communicator System. *IEEE ASRU*. 341-344.